

Sublinear Bounds for Randomized Leader Election

Shay Kutten* Gopal Pandurangan† David Peleg‡ Peter Robinson§
Amitabh Trehan *

October 18, 2012

Abstract

This paper concerns *randomized* leader election in synchronous distributed networks. A distributed leader election algorithm is presented for complete n -node networks that runs in $O(1)$ rounds and (with high probability) takes only $O(\sqrt{n} \log^{3/2} n)$ messages to elect a unique leader (with high probability). This algorithm is then extended to solve leader election on any connected non-bipartite n -node graph G in $O(\tau(G))$ time and $O(\tau(G)\sqrt{n} \log^{3/2} n)$ messages, where $\tau(G)$ is the mixing time of a random walk on G . The above result implies highly efficient (sublinear running time and messages) leader election algorithms for networks with small mixing times, such as expanders and hypercubes. In contrast, previous leader election algorithms had at least linear message complexity even in complete graphs. Moreover, super-linear message lower bounds are known for time-efficient *deterministic* leader election algorithms. Finally, an almost-tight lower bound is presented for randomized leader election, showing that $\Omega(\sqrt{n})$ messages are needed for any $O(1)$ time leader election algorithm which succeeds with high probability. It is also shown that $\Omega(n^{1/3})$ messages are needed by any leader election algorithm that succeeds with high probability, regardless of the number of the rounds. We view our results as a step towards understanding the randomized complexity of leader election in distributed networks.

1 Introduction

Background and motivation. Leader election is a classical and fundamental problem in distributed computing. It originated as the problem of regenerating the “token” in a local area *token ring* network [19] and has since then “starred” in major roles in problems across the spectrum, providing solutions for reliability by replication (or duplicate elimination), for locking, synchronization, load balancing, maintaining group

*Information Systems Group, Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology, Haifa-32000, Israel. *email:* kutten@ie.technion.ac.il, amitabh.trehan@gmail.com. Supported by the Israeli Science Foundation and by the Technion TASP center.

†Division of Mathematical Sciences, Nanyang Technological University, Singapore 637371 and Department of Computer Science, Brown University, Box 1910, Providence, RI 02912, USA. *email:* gopalpandurangan@gmail.com. Research supported in part by the following grants: Nanyang Technological University grant M58110000, Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 2 grant MOE2010-T2-2-082, and a grant from the US-Israel Binational Science Foundation (BSF).

‡Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot-76100 Israel. *email:* david.peleg@weizmann.ac.il. Supported in part by the Israel Science Foundation (grant 894/09), the United States-Israel Binational Science Foundation (grant 2008348), and the Israel Ministry of Science and Technology (infrastructures grant).

§Division of Mathematical Sciences, Nanyang Technological University, Singapore 637371. *email:* peter.robinson@ntu.edu.sg. Research supported in part by the following grants: Nanyang Technological University grant M58110000, Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 2 grant MOE2010-T2-2-082.

memberships and establishing communication primitives. As an example, the content delivery network giant Akamai uses decentralized and distributed leader election as a subroutine to tolerate machine failure and build fault tolerance in its systems [23]. In many cases, especially with the advent of large scale networks such as peer-to-peer systems [27, 28, 32], it is desirable to achieve low cost and scalable leader election, even though the guarantees may be probabilistic.

Informally, the problem of distributed leader election requires a group of processors in a distributed network to elect a unique leader among themselves, i.e., exactly one processor must output the decision that it is the leader, say, by changing a special *status* component of its state to the value *leader* [20]. All the rest of the nodes must change their status component to the value *non-leader*. These nodes need not be aware of the identity of the leader. This *implicit* variant of leader election is rather standard (cf. [20]), and is sufficient in many applications, e.g., for token generation in a token ring environment. This paper focuses on implicit leader election.

In the *explicit* variant, all the non-leaders change their status component to the value *non-leader*, and moreover, every node must also know the identity of the unique leader. This formulation may be necessary in problems where nodes coordinate and communicate through a leader, e.g., implementations of Paxos [7, 18]. In this variant, there is an obvious lower bound of $\Omega(n)$ messages (throughout, n denotes the number of nodes in the network) since every node must be informed of the leader's identity. This explicit leader election can be achieved by simply executing an (implicit) leader election algorithm and then broadcasting the leader's identity using an additional $O(n)$ messages and $O(D)$ time (where D is the diameter of the graph).

The complexity of the leader election problem and algorithms for it, especially deterministic algorithms (guaranteed to always succeed), have been well-studied. Various algorithms and lower bounds are known in different models with synchronous/asynchronous communication and in networks of varying topologies such as a cycle, a complete graph, or some arbitrary topology (e.g., see [12, 20, 24, 29, 31] and the references therein). The problem was first studied in context of a ring network by Le Lann [19] and discussed for general graphs in the influential paper of Gallager, Humblet, and Spira [8]. However, the class of complete networks has come to occupy a special position of its own and has been extensively studied [1, 10, 13, 15, 16].

The study of leader election algorithms is usually concerned with both message and time complexity. For complete graphs, Korach, Moran and Zaks [14] and Humblet [10] presented $O(n \log n)$ message algorithms. Korach, Kutten, and Moran [13] developed a general method decoupling the issue of the graph family from the design of the leader election algorithm, allowing the development of message efficient leader election algorithms for any class of graphs, given an efficient traversal algorithm for that class. When this method was applied to complete graphs, it yielded an improved (but still $\Omega(n \log n)$) message complexity. Afek and Gafni [1] presented asynchronous and synchronous algorithms, as well as a tradeoff between the message and the time complexity of synchronous *deterministic* algorithms for complete graphs in the non-simultaneous wake-up model: the results varied from a $O(1)$ -time, $O(n^2)$ -messages algorithm to a $O(\log n)$ -time, $O(n \log n)$ -messages algorithm. Singh [30] showed another trade-off that saved on time, still for algorithms with a super-linear number of messages. (Sublinear time algorithms were shown in [30] even for $O(n \log n)$ messages algorithms, and even lower times for algorithms with higher messages complexities). Afek and Gafni, as well as Korach, Moran, and Zaks [14, 16] showed a lower bound of $\Omega(n \log n)$ messages for *deterministic* algorithms in the general case. Multiple studies showed a different case where it was possible to reduce the number of messages to $O(n)$ by using a *sense of direction*—essentially, assuming some kind of a virtual ring, superimposed on the complete graph, such that the order of nodes on a ring is known to the nodes [6]. The above results demonstrate that the number of messages needed for deterministic

leader election is at least linear if nodes wake up simultaneously, or even super-linear (i.e., $\Omega(n \log n)$) if nodes are woken up by the adversary. In this paper, we focus on simultaneous wake-up.

Nevertheless, in this paper we also show that our algorithms yield sublinear message complexity even in the case where the adversary can wake up nodes at arbitrary times, which is a significant improvement over the $\Omega(n \log n)$ bound required for deterministic algorithms.

At its core, leader election is a symmetry breaking problem. For anonymous networks under some reasonable assumptions, deterministic leader election was shown to be impossible [3] (using symmetry concerns). Randomization comes to the rescue in this case; random rank assignment is often used to assign unique identifiers, as done herein. Randomization also allows us to beat the lower bounds for deterministic algorithms, albeit at the risk of a small chance of error.

Randomized asynchronous (explicit) leader election algorithms for various networks were presented by Itai and Rodeh, Scheiber and Snir, and Afek and Matias [11, 5, 2]. In particular, one of the algorithms elects a leader in a complete graph with $O(n)$ messages and $O(\log n)$ time [2]. The probability of error there tends to zero when n grows to infinity but is not given explicitly. A randomized leader election algorithm (for the explicit version) that could err with probability $O(\frac{1}{(\log n)^{\Omega(1)}})$ was presented recently in [26]¹, with time $O(\log n)$ and linear message complexity. That paper also surveys some related papers about randomized algorithms in other models that use more messages for performing leader election [9] or related tasks (e.g., probabilistic quorum systems, Malkhi et al [21]). In the context of self-stabilization, a randomized algorithm with $O(n \log n)$ messages and $O(\log n)$ time until stabilization was presented in [33].

1.1 Our Main Results

The main focus of this paper is to study how randomization can help in improving the complexity of leader election, especially message complexity in synchronous networks. We first present a (implicit) leader election algorithm for a complete network that runs in $O(1)$ time and uses only $O(\sqrt{n} \log^{3/2} n)$ messages to elect a unique leader (with high probability²). This is a significant improvement over the linear number of messages required by any deterministic algorithm (in the simultaneous wake-up model).

We also show that our algorithm works in the non-simultaneous wake-up model too, which is an even larger gap to the $\Omega(n \log n)$ message complexity bound required by any deterministic algorithm. For the explicit variant of the problem, our algorithm can be extended to use (w.h.p.) $O(n)$ messages and $O(1)$ time.

We then extend this algorithm to solve leader election on any connected (non-bipartite)³ n -node graph G in $O(\tau(G))$ time and $O(\tau(G) \sqrt{n} \log^{3/2} n)$ messages, where $\tau(G)$ is the mixing time of a random walk on G . The above result implies highly efficient (sublinear running time and messages) leader election algorithms for networks with small mixing time. In particular, for important graph classes such as expanders (used, e.g., in modeling peer-to-peer networks [4]), which have logarithmic mixing time, it implies an $O(\log n)$ time and $O(\sqrt{n} \log^{5/2} n)$ messages algorithm, and for hypercubes, which have a mixing time of $O(\log n \log \log n)$, it implies a sublinear $O(\log n \log \log n)$ time and $O(\sqrt{n} \log^{5/2} n \log \log n)$ messages algorithm.

For our algorithms, we assume that the communication is synchronous and follows the standard *CONGEST* model [25], where in each round a node can send at most one message of size $O(\log n)$ bits on a single edge. For our algorithm on general graphs, we also assume that the nodes have an estimate of the mixing time. We do not however assume that the nodes have unique IDs, hence the algorithms in this paper work also for anonymous networks. We assume that all nodes wake up simultaneously at the beginning of the execution.

¹In contrast, the probability of error in the current paper is $O(\frac{1}{n^{\Omega(1)}})$.

²Throughout, “with high probability (whp)” means with probability $\geq 1 - 1/n^{\Omega(1)}$.

³Our algorithm can be modified to work for bipartite graphs as well (cf. Section 3).

(Additional details on our distributed computing model are given later.)

Finally we show that, in general, it is not possible to improve over our algorithm substantially, by presenting an almost-tight lower bound for randomized leader election. We show that $\Omega(\sqrt{n})$ messages are needed for any $O(1)$ time leader election algorithm in a complete network which succeeds with high probability. It is also shown that $\Omega(n^{1/3})$ messages are needed by any leader election algorithm that succeeds with high probability, regardless of the number of the rounds. These lower bounds hold even in the *LOCAL* model [25], where there is no restriction on the number of bits that can be sent on each edge in each round. To the best of our knowledge, these are the first non-trivial lower bounds for randomized leader election.

1.2 Technical Contributions

The main algorithmic tool used by our randomized algorithm involves reducing the message complexity via random sampling. For general graphs, this sampling is implemented by performing random walks. Informally speaking, a small number of nodes (about $O(\log n)$), which are the candidates for leadership, initiate random walks. We show that if a sufficient number of random walks are initiated (about $\sqrt{n} \log n$), then there is a good probability that random walks originating from different candidates meet (or collide) at some node which acts as a referee. The referee notifies a winner among the colliding random walks. The algorithms use a birthday paradox type argument to show that a unique candidate node wins all competitions (i.e. is elected) with high probability. An interesting feature of that birthday paradox argument (for general graphs) is that it is applied to a setting with non-uniform selection probabilities. See Section 2 for a simple version of the algorithm that works on a complete graph. The algorithm of Section 3 is a generalization of the simple algorithm of Section 2 that works for any connected graph; however the algorithm and analysis are more involved.

The main intuition in our lower bound proof for randomized leader election is that we show that any algorithm which sends less messages than required by our lower bound has a good chance of generating runs where there are multiple potential leader candidates in the network that do not influence each other. In other words, the probability of such “disjoint” parts of the network to elect a leader is the same, which implies that there is a good probability that more than one leader is elected. Although this is conceptually easy to state, it is technically challenging to show formally since our result applies to all randomized algorithms without further restrictions.

1.3 Distributed Computing Model

The model we consider is similar to the models of [1, 10, 13, 15, 16], with the main addition of giving processors access to a private unbiased coin. Also, we do not assume unique identities. We consider a system of n nodes, represented as an undirected (not necessarily complete) graph $G = (V, E)$. Each node runs an instance of a distributed algorithm that has knowledge of n . The computation advances in synchronous rounds, where, in every round, nodes can send messages, receive messages that were sent in the same round by neighbors in G , and perform some local computation; every node has access to the outcome of unbiased private coin flips. The messages are the only means of communication; in particular, nodes cannot access the coin flips of other nodes, and do not share any memory. Throughout this paper, we assume that all nodes are awake initially and simultaneously start executing the algorithm. we discuss some relaxations of this point in a separate section.

Leader Election.

We now formally define the leader election problem. Every node u has a special variable status_u that it can set to a value in the set $\{\perp, \text{NON-ELECTED}, \text{ELECTED}\}$; initially we assume $\text{status}_u = \perp$. An algorithm A solves leader election in T rounds if, from round T on, exactly one node has its status set to ELECTED while all nodes are in state NON-ELECTED. This is the requirement for standard (implicit) leader election.

2 Randomized Leader Election in Complete Networks

To provide the intuition for our general result, let us first illustrate a simpler version of our leader election algorithm, adapted to complete networks. More specifically, this section presents an algorithm that, with high probability, solves leader election in complete networks in $O(1)$ rounds and sends no more than $O(\sqrt{n} \log^{3/2} n)$ messages. Let us first briefly describe the main ideas of Algorithm 1 (see pseudo-code below). Initially, the algorithm attempts to reduce the number of leader candidates as far as possible, while still guaranteeing that there is at least one candidate (with high probability). Non-candidate nodes enter the NON-ELECTED state immediately, and thereafter only reply to messages initiated by other nodes. Every node u becomes a candidate with probability $\frac{2 \log n}{n}$ and selects a random rank r_u chosen from some large domain. Each candidate node then randomly selects $2\lceil \sqrt{n \log n} \rceil$ other nodes as *referees* and informs all referees of its rank. The referees compute the maximum (say r_w) of all received ranks, and send a “winner” notification to the node w . If a candidate wins all competitions, i.e., receives “winner” notifications from all of its referees, it enters the ELECTED state and becomes the leader.

Algorithm 1 Randomized Leader Election in Complete Graphs

Round 1 :

- 1: Every node u decides to become a candidate with probability $\frac{2 \log n}{n}$ and generates a random rank r_u from $\{1, \dots, n^4\}$.
If a node does not become a candidate, it immediately enters the NON-ELECTED state; otherwise it executes.
- 2: **Choosing Referees:** Node u samples $2\lceil \sqrt{n \log n} \rceil$ neighbors (the *referees*) and sends a message $\langle u, r_u \rangle$ to each referee.

Round 2 :

- 3: **Winner Notification:** A referee v considers all received messages and sends a winner notification to the node w that satisfies
 $r_w \geq r_u$ for every message $\langle u, r_u \rangle$.
 - 4: **Decision:** If a node receives winner notifications from all its referees, then it enters the ELECTED state, otherwise it sets its state to NON-ELECTED.
-

Theorem 1. Consider a complete network of n nodes and assume the *CONGEST* model of communication. Algorithm 1 solves leader election with high probability, terminates in $O(1)$ rounds, and uses $O(\sqrt{n} \log^{3/2} n)$ messages with high probability.

Proof. Since all nodes enter either the elected or non-elected state after two rounds at the latest, we get the runtime bound of $O(1)$.

We now argue the message complexity bound. On expectation, there are $2 \log n$ candidate nodes. By using a standard Chernoff bound (cf. Theorem 4.4 in [22]), there are at most $7 \log n$ candidate nodes with probability at least $1 - n^{-2}$. In step 3 of the algorithm, each referee only sends messages to the candidate nodes by which it has been contacted. Since there are $O(\log n)$ candidates and each approaches $\Theta(\sqrt{n \log n})$ referees, the total number of messages sent is bounded by $O(\sqrt{n} \log^{3/2} n)$ with high probability.

Finally, we show that Algorithm 1 solves leader election with high probability. With probability $\left(1 - \frac{2 \log n}{n}\right)^n \approx \exp(-2 \log n) = n^{-2}$, no node becomes candidate. Hence the probability that at least one node is elected as leader is at least $1 - n^{-2}$. Let ℓ be the node that generates the highest random rank r_ℓ among all candidate nodes; with high probability, ℓ is unique. Clearly, node ℓ enters the ELECTED state, since it receives “winner” notifications from all its referees.

Now consider some other candidate node v . This candidate chooses its referees randomly among all nodes. Therefore, the probability that an individual referee selected by v is among the referees chosen by ℓ , is $\frac{2 \lceil \sqrt{n \log n} \rceil}{n}$. It follows that the probability that ℓ and v do not choose any common referee node is at most

$$\left(1 - 2 \sqrt{\frac{\log n}{n}}\right)^{2 \sqrt{n \log n}} \leq \exp(-4 \log n) = n^{-4},$$

which means that with high probability, some node x serves as common referee to ℓ and v . By assumption, we have $r_v < r_\ell$, which means that node v does not receive $2 \lceil \sqrt{n \log n} \rceil$ “winner” notifications, and thus it subsequently enters the NON-ELECTED state. By taking a union bound over all other candidate nodes, it follows that with probability at least $1 - \frac{1}{n}$, no other node except ℓ wins all of its competitions, and therefore, node ℓ is the only node to become a leader. \square

With a simple modification, our algorithm also works for non-simultaneous wakeup of nodes, though in that case we cannot guarantee termination in $O(1)$ rounds.

2.1 Non-Simultaneous Wake-Up of Nodes

So far, we have assumed that all nodes are up and running at the start of round 1. We now describe a simple extension of Algorithm 1 that preserves the low message complexity bound in a model where nodes are woken up at arbitrary times by the adversary (similarly to [1]). The main idea is to require a referee node v to only send winner notifications in the *first* round r when v receives a message from some candidate nodes. This ensures that the candidate u that has the highest random rank among all initially awake candidate nodes will become leader. Let R be the set of referees chosen by the winner u . To see that there is a unique leader, we observe that, analogously to the proof of Theorem 1, any candidate that wakes up in some round ($\geq r$) will choose a referee among the nodes in R with high probability. Unfortunately, we can no longer guarantee termination within $O(1)$ rounds, since the adversary can simply delay waking up all but one node u , which has only probability $\frac{2 \log n}{n}$ of becoming a candidate.

Corollary 1. *Consider a complete network of n nodes and assume the CONGEST model of communication where nodes are woken up at arbitrary times by the adversary. There is an algorithm that elects a unique leader (w.h.p.), while using $O(\sqrt{n} \log^{3/2} n)$ messages (w.h.p.).*

3 Randomized Leader Election in General Graphs

In this section, we present our main algorithm, which elects a unique leader (w.h.p.), and terminates in $O(\tau(G, n))$ rounds while using $O(\tau(G, n)\sqrt{n} \log^{3/2} n)$ messages (w.h.p.), where $\tau(G, n)$ is the mixing time of a random walk on G . Initially, a node u only knows the mixing time (or a constant factor estimate of) $\tau(G, n)$ (defined below in (1)); in particular u does not have any a priori knowledge about the actual topology of G .

The algorithm presented here requires nodes to perform random walks on the network by token forwarding in order to choose sufficiently many referee nodes at random. Thus essentially random walks perform the role of sampling as done in Algorithm 1 and is conceptually similar. Whereas in the complete graph randomly chosen nodes act as referees, here any intermediate node (in the random walk) that sees tokens from two competing candidates can act as a referee and notify the winner. One slight complication we have to deal with in the general setting is that in the *CONGEST* model it is impossible to perform too many walks in parallel along an edge. We solve this issue by sending only the *count* of tokens that need to be sent by a particular candidate, and not the tokens themselves.

While using random walks can be viewed as a generalization of the sampling performed in Algorithm 1, showing that two candidate nodes intersect in at least one referee leads to an interesting balls-into-bins scenario where balls (i.e., random walks) have a *non-uniform* probability to be placed in some bin (i.e., reach a referee node). This non-uniformity of the random walk distribution stems from the fact that G might not be a regular graph. We show that the non-uniform case does not worsen the probability of two candidates reaching a common referee, and hence an analysis similar to the one given for complete graphs goes through.

We now introduce some basic notation for random walks. Suppose that $V = \{u_1, \dots, u_n\}$ and let d_i denote the degree of node i . The $n \times n$ *transition matrix* \mathbf{A} of G has entries $a_{i,j} = \frac{1}{d_i}$ if there is an edge $(i, j) \in E$, otherwise $a_{i,j} = 0$. Entry $a_{i,j}$ gives the probability that a random walk moves from node u_i to node u_j . The position of a random walk after k steps is represented by a probability distribution π_k determined by \mathbf{A} . If some node u_i starts a random walk, the initial distribution π_0 of the walk is an n -dimensional vector having all zeros except at index i where it is 1. Once node u has chosen a random neighbor to forward the token, the distribution of the walk after 1 step is given by $\pi_1 = \mathbf{A}\pi_0$ and in general we have $\pi_k = \mathbf{A}^k\pi_0$. If G is non-bipartite and connected, then the distribution of the walk will eventually converge to the *stationary distribution* $\pi_* = (b_1, \dots, b_n)$, which has entries $b_i = \frac{d_i}{2|E|}$ and satisfies $\pi_* = \mathbf{A}\pi_*$.

We define the *mixing time* $\tau(G, n)$ of a graph G with n nodes as the minimum k such that, for all starting distributions π_0 ,

$$\|\mathbf{A}^k\pi_0 - \pi_*\|_\infty \leq \frac{1}{2n}, \quad (1)$$

where $\|\cdot\|_\infty$ denotes the usual maximum norm on a vector. Clearly, if G is a complete network, then $\tau(G, n) = 1$. For expander graphs it is well known that $\tau(G, n) \in O(\log n)$. Note that mixing time is well-defined only for non-bipartite graphs; however, by using a lazy random walk strategy (i.e., with probability $1/2$ stay at the current node; otherwise proceed as usual) our algorithm will work for bipartite graphs as well.

First, we prove a useful lemma:

Lemma 1. *Consider p balls that are placed into n bins according to some probability distribution π and let p_i be the i -th entry of π . Let X_i be the indicator random variable that is 1 if there is a collision (of random walks) at referee node i . Then $\mathbf{P}[\bigcap_{i=1}^n (X_i = 0)]$ is maximized for the uniform distribution.*

Algorithm 2 Randomized Leader Election

- 1: **VAR** `origin` $\leftarrow 0$; `winner-so-far` $\leftarrow \perp$
- 2: Initially, node u decides to become a candidate with probability $\frac{2 \log n}{n}$ and generates a random rank r_u from $\{1, \dots, n^4\}$.

Initiating Random Walks:

- 3: Node u creates $2\lceil \sqrt{n \log n} \rceil$ tokens of type $\langle r_u, k \rangle$.
- 4: Node u starts $2\lceil \sqrt{n \log n} \rceil$ random walks (called *competitions*), each of which is represented by the random walk token $\langle r_u, k \rangle$ (of $O(\log n)$ bits) where r_u represents u 's random rank. The counter k is the number (initially 1) of walks that are represented by this token (explained in Line 8).

Disqualifying hopeless candidates (note that any node can be a referee and notify winner/loser):

- 5: A node v discards every received token $\langle r_u, k \rangle$ if v has received (possibly in the same round) a token r_w with $r_w > r_u$.
- 6: **if** a received token $\langle r_w, k' \rangle$ is not discarded and `winner-so-far` $\neq r_w$ **then**
- 7: Node v remembers the port of an arbitrarily chosen neighbor that sent one of the (possibly merged) tokens containing r_w in its variable `origin` and sets its variable `winner-so-far` to r_w .

Token Forwarding:

- 8: Let $\mu = \langle r_u, k \rangle$ be a token received by v and suppose that μ is not discarded in Line 5. For simplicity, we consider all distinct tokens that arrive in the current round containing the same value r_u at v to be merged into a single token $\langle r_u, k \rangle$ before processing where k holds the accumulated count. Node v randomly samples k times from its neighbors. If a neighbor x was chosen $k_x \leq k$ times, v sends a token $\langle r_u, k_x \rangle$ to x .

Notifying a Winner in round $\tau(G, n)$:

- 9: **if** `winner-so-far` $\neq \perp$ **then**
- 10: Suppose that node v has not discarded some token generated by w . According to Line 5, w has generated the largest rank among all tokens seen by v .
- 11: Node v generates a winner notification $\langle \text{WIN}, r_w, cnt \rangle$ for r_w and sends it to the neighbor stored in `origin` (cf. Line 7). The field `cnt` is set to 1 by v and contains the number of winner notifications represented by this token.
- 12: If a node u receives (possibly) multiple winner notifications for r_w , it simply forwards a token $\langle \text{WIN}, r_w, cnt' \rangle$ to the neighbor stored in `origin` where cnt' is the accumulated count of all received tokens.

Decision:

- 13: If a node wins all competitions, i.e., receives $2\lceil \sqrt{n \log n} \rceil$ winner notifications it enters the ELECTED state, otherwise it sets its state to NON-ELECTED.
-

Proof. By definition, we have

$$\mathbb{P}[X_i = 1] = (1 - (1 - p_i)^\rho)^2.$$

Note that the events $X_i = 1$ and $X_j = 1$ are not necessarily independent. A common technique to treat dependencies in balls-into-bins scenarios is the Poisson approximation where we consider the number of balls in each bin to be independent Poisson random variables with mean ρ/n . This means we can apply Corollary 5.11 of [22], which states that if some event E occurs with probability p in the Poisson case, it occurs with probability at most $2p$ in the exact case, i.e., we only lose a constant factor by using the Poisson approximation. A precondition for applying Corollary 5.11, is that the probability for event E

monotonically decreases (or increases) in the number of balls, which is clearly the case when counting the number of collisions of balls. Considering the Poisson case, we get

$$\begin{aligned} \mathbf{P} \left[\bigcap_{i=1}^n (X_i = 0) \right] &= \prod_{i=1}^n \mathbf{P}[X_i = 0] = \prod_{i=1}^n (1 - (1 - (1 - p_i)^\rho)^2) . \\ &\leq \prod_{i=1}^n (1 - (1 - e^{-p_i \rho})^2) \leq \prod_{i=1}^n (1 - (p_i \rho)^2) \\ &\leq \prod_{i=1}^n e^{-p_i^2 \rho^2} = \exp \left(-\rho^2 \sum_{i=1}^n p_i^2 \right) . \end{aligned}$$

To maximize $\mathbf{P} [\bigcap_{i=1}^n (X_i = 0)]$, it is thus sufficient to minimize $\sum_{i=1}^n p_i^2$ under the constraint $\sum_{i=1}^n p_i = 1$. Using Lagrangian optimization it follows that this is minimized for the uniform distribution. \square

Theorem 2. *Consider a non-bipartite network G of n nodes with mixing time $\tau(G, n)$, and assume the *CONGEST* model of communication. Algorithm 2 solves leader election with high probability, terminates within $O(\tau(G, n))$ rounds, and uses $O(\tau(G, n)\sqrt{n} \log^{3/2} n)$ messages with high probability.*

Proof. We first argue the message complexity bound. On expectation, there are $\Theta(\log n)$ candidate nodes. By using a standard Chernoff bound (cf. Theorem 4.4 in [22]), there are at most $7 \log n$ candidate nodes with probability at least $1 - n^{-2}$. Every candidate node u contacts $\Theta(\sqrt{n \log n})$ referee nodes and initiates a random walk of length $\tau(G, n)$, for each of the $\Theta(\sqrt{n \log n})$ referees. By the description of the algorithm, each referee node only sends messages to the candidate nodes by which it has been contacted. Since we have $O(\log n)$ candidates, the total number of messages sent is bounded by $O(\tau(G, n)\sqrt{n} \log^{3/2} n)$ with high probability.

The running time bound depends on the time that it takes to complete the $2\lceil \sqrt{n \log n} \rceil$ random walks in parallel and the notification of the winner. By Line 5, it follows that a node only forwards at most one token to any neighbor in a round, thus there is no delay due to congestion. Moreover, for notifying the winner, nodes forward the winner notification for winner w to the neighbor stored in `origin`. According to Line 7, a node sets `origin` to a neighbor from which it has received the first token originated from w . Thus there can be no loops when forwarding the winner notifications, which reach the winner w in at most $\tau(G, n)$ rounds.

We now argue that Algorithm 2 solves leader election with high probability. Similarly to Algorithm 1, it follows that there will be at least one leader with high probability. Let ℓ be the candidate that generated the (unique) highest random rank among all candidates and consider some other candidate node v , i.e., we have that $r_v < r_\ell$ by assumption. By the description of the algorithm, node v chooses its referees by performing $\rho = 2\lceil \sqrt{n \log n} \rceil$ random walks of length $\tau(G, n)$. We cannot argue the same way as in the proof of Algorithm 1, since in general, the stationary distribution of G might not be the uniform distribution vector $(\frac{1}{n}, \dots, \frac{1}{n})$. Let p_i be the i -th entry of the stationary distribution. Let X_i be the indicator random variable that is 1 if there is a collision (of random walks) at referee node i . We have $\mathbf{P}[X_i = 1] = (1 - (1 - p_i)^\rho)^2$. We want to show that the probability of error (i.e., having no collisions) is small; in other words, we want to upper bound $\mathbf{P} [\bigcap_{i=1}^n (X_i = 0)]$. Lemma 1 shows that it is sufficient to obtain a bound for the case when the stationary distribution is uniform. By (1), the probability of such a walk hitting any of the referees chosen by ℓ , is at least $\frac{2\sqrt{n \log n}}{2n}$. It follows that the probability that ℓ and v do not choose a common referee node is

at most

$$\left(1 - \sqrt{\frac{\log n}{n}}\right)^{2\sqrt{n \log n}} \leq \exp(-2 \log n). \quad (2)$$

Therefore, the event that node v does not receive sufficiently many winner notifications, happens with probability $\geq 1 - n^{-2}$, which requires v to enter the NON-ELECTED state. By taking a union bound over all other candidate nodes, it follows that with high probability no other node except ℓ will win all of its competitions, and therefore, node ℓ is the only node to become a leader with probability at least $1 - \frac{1}{n}$. \square

4 Lower Bound

In this section we prove a lower bound on the number of messages required by any algorithm that solves leader election with probability at least $1 - 1/n$.

Our model assumes that all processors execute the exact same algorithm and have access to an unbiased private coin. So far we have assumed that nodes are *not* equipped with unique ids. Nevertheless, our lower bound still holds even if the nodes start with unique ids.

Our lower bound applies to all algorithms that send only $o(\sqrt{n})$ messages with probability at least $1 - 1/n$. In other words, the result still holds for algorithms that have small but nonzero probability for producing runs where the number of messages sent is much larger (i.e., $\Omega(\sqrt{n})$). We show the result for the \mathcal{LOCAL} model, which implies the same for the $\mathcal{CONGEST}$ model.

Theorem 3. *Consider any algorithm A that uses $f(n)$ messages (of arbitrary size) with high probability on a complete network of n nodes. If A solves leader election in $O(1)$ rounds with high probability, then $f(n) \in \Omega(\sqrt{n})$. Moreover, $f(n) \in \Omega(n^{1/3})$ for any algorithm A using any number of rounds that solves leader election with high probability. This holds even if nodes are equipped with unique identifiers (chosen by the adversary).*

Proof. We first show the result for the case where nodes are anonymous, i.e., are *not* equipped with unique identifiers, and later on extend the impossibility to the non-anonymous case by an easy reduction.

Assume that there is some algorithm A that solves leader election with high probability but sends only $f(n)$ messages. The remainder of the proof involves showing that this yields a contradiction. Consider a complete network where for every node, the adversary chooses the connections of its ports as a random permutation on $\{1, \dots, n - 1\}$.

For a given run α of an algorithm, define the *communication graph* $\mathcal{C}^r(\alpha)$ to be a directed graph on the given set of n nodes where there is an edge from u to v if and only if u sends a message to v in some round $r' \leq r$ of the run α . For any node u , denote the *state of u in round r of the run α* by $\sigma_r(u, \alpha)$. Let Σ be the set of all node states possible in algorithm A . (When α is known, we may simply write \mathcal{C}^r and $\sigma_r(u)$.) With each node $u \in \mathcal{C}^r$, associate its state $\sigma_r(u)$ in \mathcal{C}^r , the communication graph of round r . We say that node u *influences node w by round r* if there is a directed path from u to w in \mathcal{C}^r . (Our notion of influence is more general than the causality based “happens-before” relation of [17], since a directed path from u to w is necessary but not sufficient for w to be causally influenced by u .) A node u is an *initiator* if it is not influenced before sending its first message. Note that a mute node that never receives any messages is also an initiator. For every initiator u , we define the *influence cloud* \mathcal{IC}_u^r as the pair $\mathcal{IC}_u^r = (\mathcal{C}_u^r, S_u^r)$, where $\mathcal{C}_u^r = \langle u, w_1, \dots, w_k \rangle$ is the ordered set of all nodes that are influenced by u , namely, that are reachable along a directed path in \mathcal{C}^r from u , ordered by the time by which they joined the cloud, and $S_u^r = \langle \sigma_r(u, \alpha), \sigma_r(w_1, \alpha), \dots, \sigma_r(w_k, \alpha) \rangle$ is their configuration after round r , namely, their

current tuple of states. (In what follows, we sometimes abuse notation by referring to the ordered node set C_u^r as the influence cloud of u .) Note that a *passive* (non-initiator) node v does not send any messages before receiving the first message from some other node.

Since we are only interested in algorithms that send a finite number of messages, in every execution α there is some round $\rho = \rho(\alpha)$ by which no more messages are sent.

In general, it is possible that in a given execution, two influence clouds $C_{u_1}^r$ and $C_{u_2}^r$ intersect each other over some common node v , if v happens to be influenced by both u_1 and u_2 . The following lemma shows that the low message complexity of algorithm A yields a good probability for all influence clouds to be disjoint from each other.

Hereafter, we fix a run α of algorithm A . Let \hat{N}_i be the event that there is no intersection between (the node sets of) the influence clouds existing at the end of round i , i.e., $C_u^i \cap C_{u'}^i = \emptyset$ for every two initiators u, u' . Let $N_r = \bigwedge_{i=1}^r \hat{N}_i$. Let $N = N_\rho$ be the corresponding event at the end of the run α . Let M be the event that algorithm A sends no more than $f(n)$ messages in the run α .

Lemma 2. *Assume that $\mathbf{P}[M] \geq 1 - \frac{1}{n}$. Then either of the following two conditions is sufficient to ensure that $\mathbf{P}[N \wedge M] \geq 1 - o(1)$:*

- (a) $f(n) \in o(\sqrt{n})$ and A terminates in $O(1)$ rounds, or
- (b) $f(n) \in o(n^{1/3})$ and A terminates (in an arbitrary number of rounds).

Proof. Under the assumption that $\mathbf{P}[M] \geq 1 - \frac{1}{n}$, and since $\mathbf{P}[N \wedge M] = \mathbf{P}[N | M] \cdot \mathbf{P}[M]$, it suffices to show that $\mathbf{P}[N | M] \geq 1 - o(1)$ under the assumptions (a) or (b).

We first show the claim assuming (a). To prove the claim, we show by induction on r that $\mathbf{P}[N_r \wedge M] \geq 1 - o(1)$ for every $0 \leq r \leq \rho$. For $r = 0$ the claim is immediate. Now assume the claim for rounds up to $r - 1$ and consider round r . Consider some cloud C^r and any node $v \in C^r$. Conditioning on M , there are at most $f(n)$ nodes in all other clouds except C^r . Recall that the port numbering of every node was chosen uniformly at random and, since we assume M , any node knows the destinations of at most $f(n)$ of its ports in any round. To send a message to a node in another cloud, v must hit upon one of the (at most $f(n)$) ports leading to other clouds, from among its (at least $n - f(n)$) yet unexposed ports. Therefore, the probability that a message sent by v reaches a node in another cloud is at most $\frac{f(n)}{n - f(n)}$, which implies that

$$\begin{aligned} \mathbf{P}[\hat{N}_r | N_{r-1} \wedge M] &\geq \left(1 - \frac{f(n)}{n - f(n)}\right)^{f(n)} \geq \exp\left(-\frac{2f^2(n)}{n - f(n)}\right) \\ &\geq 1 - \frac{2f^2(n)}{n - f(n)} = 1 - o(1). \end{aligned} \tag{3}$$

The probability that there are no intersections up to round r satisfies

$$\mathbf{P}[N_r | M] = \mathbf{P}[\hat{N}_r \wedge N_{r-1} | M] = \mathbf{P}[\hat{N}_r | N_{r-1} \wedge M] \cdot \mathbf{P}[N_{r-1} \wedge M] = 1 - o(1),$$

where the last equality follows by Eq. (3) and the inductive hypothesis. This completes the inductive proof and establishes that $\mathbf{P}[N \wedge M] = \mathbf{P}[N_\rho \wedge M] = 1 - o(1)$.

Next, we prove the claim assuming (b), i.e., $f(n) \in o(n^{1/3})$. Let D_v be the event that node v sends a message to some other cloud on some round of the run α . By the same argument as in case (a), the probability that a message sent by v reaches a node in another cloud is at most $\frac{f(n)}{n - f(n)}$. Moreover, v sends at most $f(n)$ messages. Therefore, $\mathbf{P}[D_v | M] \leq \frac{f^2(n)}{n - f(n)}$. Note that \bar{N} , the complementary event to N , satisfies $\bar{N} = \bigvee_{i=1}^{f(n)} D_{v_i}$, and, considering that at most $f(n)$ nodes can send a message in total, we have that

$\mathbf{P}[\bar{N} \mid M] \leq \frac{f^2(n)f(n)}{n-f(n)} = \frac{f^3(n)}{n-f(n)}$. Clearly, this probability is $o(1)$, since by assumption $f(n) \in o(n^{1/3})$, thus we get $\mathbf{P}[N \wedge M] = 1 - o(1)$. \square

We next consider *potential cloud configurations*, namely, $Z = \langle \sigma_0, \sigma_1, \dots, \sigma_k \rangle$, where $\sigma_i \in \Sigma$ for every i , and more generally, *potential cloud configuration sequences* $\bar{Z}^r = (Z^1, \dots, Z^r)$, where each Z^i is a potential cloud configuration, which may potentially occur as the configuration tuple of some influence clouds in round i of some execution of Algorithm A (in particular, the lengths of the cloud configurations Z^i are monotonely non-decreasing). We study the occurrence probability of potential cloud configuration sequences.

We say that the potential cloud configuration $Z = \langle \sigma_0, \sigma_1, \dots, \sigma_k \rangle$ is *realized* by the initiator u in round r of execution α if the influence cloud $\mathcal{IC}_u^r = (C_u^r, S_u^r)$ has the same node states in S_u^r as those of Z , or more formally, $S_u^r = \langle \sigma_r(u, \alpha), \sigma_r(w_1, \alpha), \dots, \sigma_r(w_k, \alpha) \rangle$, such that $\sigma_r(u, \alpha) = \sigma_0$ and $\sigma_r(w_i, \alpha) = \sigma_i$ for every $i \in [1..k]$. In this case, the influence cloud \mathcal{IC}_u^r is referred to as a *realization* of the potential cloud configuration Z . (Note that a potential cloud configuration may have many different realizations.)

More generally, we say that the potential cloud configuration sequence $\bar{Z}^r = (Z^1, \dots, Z^r)$ is realized by the initiator u in execution α if for every round $i = 1, \dots, r$, the influence cloud \mathcal{IC}_u^i is a realization of the potential cloud configuration Z^i . In this case, the sequence of influence clouds of u up to round r , $\bar{\mathcal{IC}}_u^r = \langle \mathcal{IC}_u^1, \dots, \mathcal{IC}_u^r \rangle$, is referred to as a realization of \bar{Z}^r . (Again, a potential cloud configuration sequence may have many different realizations.)

For a potential cloud configuration Z , let $E_u^r(Z)$ be the event that Z is realized by the initiator u in (round r of) the run of algorithm A . For a potential cloud configuration sequence \bar{Z}^r , let $E_u(\bar{Z}^r)$ denote the event that \bar{Z}^r is realized by the initiator u in (the first r rounds of) the run of algorithm A .

Lemma 3. *Restrict attention to executions of algorithm A that satisfy event N , namely, in which all stable influence clouds are disjoint. Then $\mathbf{P}[E_u(\bar{Z}^r)] = \mathbf{P}[E_v(\bar{Z}^r)]$ for every $r \in [1, \rho]$, every potential cloud configuration sequence \bar{Z}^r , and every two initiators u and v .*

Proof. The proof is by induction on r . Initially, in round 1, all possible influence clouds of algorithm A are singletons, i.e., their node sets contain just the initiator. Neither u nor v have received any messages from other nodes. This means that $\mathbf{P}[\sigma_1(u) = s] = \mathbf{P}[\sigma_1(v) = s]$ for all $s \in \Sigma$, thus any potential cloud configuration $Z^1 = \langle s \rangle$ has the same probability of occurring for any initiator, implying the claim.

Assuming that the result holds for round $r - 1 \geq 1$, we show that it still holds for round r . Consider a potential cloud configuration sequence $\bar{Z}^r = (Z^1, \dots, Z^r)$ and two initiators u and v . We need to show that \bar{Z}^r is equally likely to be realized by u and v , conditioned on the event N . By the inductive hypothesis, the prefix $\bar{Z}^{r-1} = (Z^1, \dots, Z^{r-1})$ satisfies the claim. Hence it suffices to prove the following. Let p_u be the probability of the event $E_u^r(Z^r)$ conditioned on the event $N \wedge E_u(\bar{Z}^{r-1})$. Define the probability p_v similarly for v . Then it remains to prove that $p_u = p_v$.

To do that we need to show, for any state $\sigma_j \in Z^r$, that the probability that $w_{u,j}$, the j th node in \mathcal{IC}_u^r , is in state σ_j , conditioned on the event $N \wedge E_u(\bar{Z}^{r-1})$, is the same as the probability that $w_{v,j}$, the j th node in \mathcal{IC}_v^r , is in state σ_j , conditioned on the event $N \wedge E_v(\bar{Z}^{r-1})$.

There are two cases to be considered. The first is that the potential influence cloud Z^{r-1} has j or more states. Then by our assumption that events $E_u(\bar{Z}^{r-1})$ and $E_v(\bar{Z}^{r-1})$ hold, the nodes $w_{u,j}$ and $w_{v,j}$ were already in u 's and v 's influence clouds, respectively, at the end of round $r - 1$. The node $w_{u,j}$ changes its state from its previous state, σ_j' , to σ_j on round r as the result of receiving some messages M_1, \dots, M_ℓ from neighbors x_1^u, \dots, x_ℓ^u in u 's influence cloud \mathcal{IC}_u^{r-1} , respectively. In turn, node x_j^u sends message M_j to $w_{u,j}$ on round r as the result of being in a certain state $\sigma_r(x_j^u)$ at the beginning of round r (or equivalently, on

the end of round $r - 1$ and making a certain random choice (with a certain probability q_j for sending M_j to $w_{u,j}$). But if one assumes that the event $E_v(\bar{Z}^{r-1})$ holds, namely, that \bar{Z}^{r-1} is realized by the initiator v , then the corresponding nodes x_1^v, \dots, x_ℓ^v in v 's influence cloud \mathcal{IC}_v^{r-1} will be in the same respective states ($\sigma_r(x_j^v) = \sigma_r(x_j^u)$ for every j) on the end of round $r - 1$, and therefore will send the messages M_1, \dots, M_ℓ to the node $w_{v,j}$ with the same probabilities q_j . Also, on the end of round $r - 1$, the node $w_{v,j}$ is in the same state σ'_j as $w_{u,j}$ (assuming event $E_v(\bar{Z}^{r-1})$). It follows that the node $w_{v,j}$ changes its state to σ_j on round r with the same probability as the node $w_{u,j}$.

The second case to be considered is when the potential influence cloud Z^{r-1} has fewer than j states. This means (conditioned on the events $E_u(\bar{Z}^{r-1})$ and $E_v(\bar{Z}^{r-1})$ respectively) that the nodes $w_{u,j}$ and $w_{v,j}$ were not in the respective influence clouds on the end of round $r - 1$. Rather, they were both passive nodes. By an argument similar to that made for round 1, any pair of (so far) passive nodes have equal probability of being in any state. Hence $\mathbf{P}[\sigma_{r-1}(w_{u,j}) = s] = \mathbf{P}[\sigma_{r-1}(w_{v,j}) = s]$ for all $s \in \Sigma$. As in the former case, the node $w_{u,j}$ changes its state from its previous state, σ'_j , to σ_j on round r as the result of receiving some messages M_1, \dots, M_ℓ from neighbors x_1^u, \dots, x_ℓ^u that are already in u 's influence cloud \mathcal{IC}_u^{r-1} , respectively. By a similar analysis, it follows that the node $w_{v,j}$ changes its state to σ_j on round r with the same probability as the node $w_{u,j}$. \square

We now conclude that for every potential cloud configuration Z , every execution α and every two initiators u and v , the events $E_u^\rho(Z)$ and $E_v^\rho(Z)$ are equally likely. More specifically, we say that the potential cloud configuration Z is *equi-probable for initiators u and v* if $\mathbf{P}[E_u^\rho(Z) \mid N] = \mathbf{P}[E_v^\rho(Z) \mid N]$. Although a potential cloud configuration Z may be the end-cloud of many different potential cloud configuration sequences, and each such potential cloud configuration sequence may have many different realizations, the above lemma implies the following (integrating over all possible choices).

Corollary 2. *Restrict attention to executions of algorithm A that satisfy event N , namely, in which all (final) stable influence clouds are disjoint. Consider two initiators u and v and a potential cloud configuration Z . Then Z is equi-probable for u and v .*

By assumption, algorithm A errs with probability $p_{\text{err}} \leq 1/n$. Let S be the event that A elects exactly one leader. We get

$$\mathbf{P}[S \mid M \wedge N] \geq \mathbf{P}[M \wedge N] - p_{\text{err}} = 1 - o(1). \quad (4)$$

Conditioning on event $M \wedge N$, let X be the random variable that represents the number of disjoint influence clouds generated by algorithm A . By Cor. 2, each of the initiators has the same probability p of generating a leader cloud. Algorithm A succeeds whenever event S occurs. Its success probability assuming $X = c$ is

$$\mathbf{P}[S \mid M \wedge N \wedge (X = c)] = cp(1 - p)^{c-1}. \quad (5)$$

For any given c , the value of (5) is maximized if $p = \frac{1}{c}$, which yields that $\mathbf{P}[S \mid M \wedge N \wedge (X = c)] \leq 1/e$ for any c . It follows that $\mathbf{P}[S \mid M \wedge N] \leq 1/e$ as well. This, however, is a contradiction to (4) and completes the proof of Theorem 3 for algorithms without unique identifiers.

We now briefly argue why our result holds for any algorithm B that runs in a model where nodes are equipped with unique ids (chosen by the adversary). Suppose that, w.h.p., B succeeds in electing a leader while sending only $f(n)$ messages. Now consider an algorithm B' in our model that is identical to B with the difference that before performing any other computation, every node generates a random number from the range $[1, \dots, n^4]$ and uses this value instead of the unique id. Let I be the event that all node ids are distinct; clearly I happens with high probability. Therefore, by the success probability of B , it follows that

B' also succeeds with probability $1 - o(1)$ (conditioned on I), which contradicts our result for algorithms without unique ids. This completes the proof of Theorem 3. \square

5 Conclusion

We studied the role played by randomization in distributed leader election. Some open questions on randomized leader election are raised by our work: (1) Can we find (universal) upper and lower bounds for general graphs? (2) Is $\Omega(\sqrt{n})$ a lower bound on the number messages needed for a complete graph, *regardless* of the number of rounds?

References

- [1] Y. Afek and E. Gafni. Time and message bounds for election in synchronous and asynchronous complete networks. *SICOMP*, 20(2):376–394, 1991.
- [2] Yehuda Afek and Yossi Matias. Elections in anonymous networks. *Inf. Comput.*, 113(2):312–330, 1994.
- [3] Dana Angluin. Local and global properties in networks of processors (extended abstract). In *STOC*, pages 82–93, 1980.
- [4] John Augustine, Gopal Pandurangan, Peter Robinson, and Eli Upfal. Towards robust and efficient distributed computation in dynamic peer-to-peer networks. In *SODA*, 2012.
- [5] Marc Snir Baruch Scheiber. Calling names on nameless networks. *Inf. Comput.*, 113(1):80–101, 1994.
- [6] Loui M. C., Matsushita T. A., and West D. B. Election in a complete network with a sense of direction. *Information Processing Letters*, 22(4):185–187, 1986.
- [7] Tushar Deepak Chandra, Robert Griesemer, and Joshua Redstone. Paxos made live - an engineering perspective (2006 invited talk). In *Proceedings of the 26th Annual ACM Symposium on Principles of Distributed Computing*, 2007.
- [8] R. G. Gallager, P. A. Humblet, and P. M. Spira. A distributed algorithm for minimum-weight spanning trees. *ACM Trans. Program. Lang. Syst.*, 5(1):66–77, January 1983.
- [9] Indranil Gupta, Robbert van Renesse, and Kenneth P. Birman. A probabilistically correct leader election protocol for large groups. In *Proceedings of the 14th International Conference on Distributed Computing*, DISC '00, pages 89–103, 2000.
- [10] P Humblet. Electing a leader in a clique in $O(n \log n)$ messages. Intern. Memo., Laboratory for Information and Decision Systems, M.I.T., Cambridge, Mass, 1984.
- [11] Alon Itai and Michael Rodeh. Symmetry breaking in distributed networks. *Inf. Comput.*, 88(1):60–87, 1990.
- [12] Maleq Khan, Fabian Kuhn, Dahlia Malkhi, Gopal Pandurangan, and Kunal Talwar. Efficient distributed approximation algorithms via probabilistic tree embeddings. In *Proceedings of the twenty-seventh ACM symposium on Principles of distributed computing*, PODC '08, pages 263–272, New York, NY, USA, 2008. ACM.

- [13] E. Korach, S. Kutten, and S. Moran. A modular technique for the design of efficient distributed leader finding algorithms. *ACM Trans. Program. Lang. Syst.*, 12(1):84–101, January 1990.
- [14] E. Korach, S. Moran, and S. Zaks. Tight lower and upper bounds for some distributed algorithms for a complete network of processors. In *PODC 1984*, pages 199–207, New York, NY, USA, 1984. ACM.
- [15] E. Korach, S. Moran, and S. Zaks. The optimality of distributive constructions of minimum weight and degree restricted spanning trees in a complete network of processors. *SIAM Journal on Computing*, 16(2):231–236, 1987.
- [16] E. Korach, S. Moran, and S. Zaks. Optimal lower bounds for some distributed algorithms for a complete network of processors. *Theoretical Computer Science*, 64(1):125 – 132, 1989.
- [17] Leslie Lamport. Time, clocks, and the ordering of events in a distributed system. *Commun. ACM*, 21(7):558–565, 1978.
- [18] Leslie Lamport. The part-time parliament. *ACM Trans. Comput. Syst.*, 16(2):133–169, May 1998.
- [19] Gérard Le Lann. Distributed systems - towards a formal approach. In *IFIP Congress*, pages 155–160, 1977.
- [20] Nancy Lynch. *Distributed Algorithms*. Morgan Kaufman Publishers, Inc., San Francisco, USA, 1996.
- [21] Dahlia Malkhi, Michael Reiter, and Rebecca Wright. Probabilistic quorum systems. In *PODC 1997*, pages 267–273, New York, NY, USA, 1997. ACM.
- [22] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2004.
- [23] Erik Nygren, Ramesh K. Sitaraman, and Jennifer Sun. The akamai network: a platform for high-performance internet applications. *SIGOPS Oper. Syst. Rev.*, 44(3):2–19, August 2010.
- [24] David Peleg. Time-optimal leader election in general networks. *Journal of Parallel and Distributed Computing*, 8(1):96 – 99, 1990.
- [25] David Peleg. *Distributed Computing: A Locality-Sensitive Approach*. SIAM, 2000.
- [26] Murali Krishna Ramanathan, Ronaldo A. Ferreira, Suresh Jagannathan, Ananth Grama, and Wojciech Szpankowski. Randomized leader election. *Distributed Computing*, pages 403–418, 2007.
- [27] Sylvia Ratnasamy, Paul Francis, Mark Handley, Richard Karp, and Scott Shenker. A scalable content-addressable network. In *SIGCOMM 2001*, pages 161–172, New York, NY, USA, 2001. ACM.
- [28] Antony I. T. Rowstron and Peter Druschel. Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems. In *Proceedings of the IFIP/ACM International Conference on Distributed Systems Platforms Heidelberg*, Middleware ’01, pages 329–350. Springer-Verlag, 2001.
- [29] Nicola Santoro. *Design and Analysis of Distributed Algorithms (Wiley Series on Parallel and Distributed Computing)*. Wiley-Interscience, 2006.
- [30] G. Singh. Efficient distributed algorithms for leader election in complete networks. In *ICDCS*, pages 472–479, 1991.

- [31] Gerard Tel. *Introduction to distributed algorithms*. Cambridge University Press, New York, NY, USA, 1994.
- [32] B.Y. Zhao, Ling Huang, J. Stribling, S.C. Rhea, A.D. Joseph, and J.D. Kubiatowicz. Tapestry: a resilient global-scale overlay for service deployment. *Selected Areas in Communications, IEEE Journal on*, 22(1):41 – 53, jan. 2004.
- [33] Dmitry Zinenko and Shay Kutten. Low communication self-stabilization through randomization. In *DISC*, 2010.